Understanding statistical variation: A response to Sharma

Jim Farmer
Macquarie University
<jafarmer@laurel.ocs.mq.edu.au>

Issue 21(2) of the *Australian Senior Mathematics Journal* contains the paper "Exploring pre-service teachers' understanding of statistical variation: Implications for teaching and research" by Sashi Sharma.

Sharma describes a study "designed to investigate pre-service teachers' acknowledgment of variation in sampling and distribution environments." In the study, "24 pre-service teacher education students completed a questionnaire about variation during one of the tutorials."

The second of the two questions used in the study is reproduced below.

(a) Imagine you threw a die 60 times. Fill in the table below to show how many time each number might come up.

Number on die	How many times it might come up?
1	
2	
3	
4	
5	
6	
TOTAL	60

(b) Why do you think these numbers are reasonable?"

My own response would have been to interpret the question literally and write "0, 1, 2, … 60" in each of the six empty cells, since this is the complete list of the number of times that each number could appear. However, this says more about my tendency to try to mess with the minds of pollsters than it does about my understanding of statistical variation. Based on the discussion in Sharma it seems that none of the students surveyed took this aberrant approach, but rather provided one number for each empty cell in the table, so we can move on to more serious issues.

Sharma provides some student responses. Here are two:

"12, 11, 9, 10, 8, 9 — because they are around the expected, but you can't really tell."

"10, 10, 10, 10, 10 because each number has the same chance of being rolled."

The first student response quoted above was labelled "Statistical" on the grounds that it demonstrated variation about the mean. The second response was labelled "Partial-statistical" and was regarded as showing less understanding, since it contained no variation about the mean.

Sharma states: "Students' numerical responses on Item 2 were coded on two scales, a centring scale (10, 10, 10, 10, 10, 10) and a scale for variation (low, appropriate, high)."

The centring scale is not explained further. Perhaps any response where the 6 numbers total to 60 can be viewed as appropriately centred. However, this highlights the deficiency of the first student response: it cannot actually occur since the numbers given total 59 rather than 60.

Perhaps this was a typographical error. Let us replace the given response by 12, 11, 9, 11, 8, 9, which does sum to 60, and try continue the argument.

Sharma notes that "the open-ended nature of the questions and the lack of guidance given to students regarding what was required of them certainly influenced how students explained their understanding. Some of the conceptions addressed in this paper may actually be due to misinterpretation of the questions."

It seems that the survey did not indicate that its purpose was to assess understanding of variation. When those surveyed have no clearly identified purpose for their response, they are free to impose their own perceived purpose on the question. I would like to elaborate on some possible alternative interpretations of the die roll question.

There is nothing in the question requiring any assessment of probability, so it seems valid to interpret the question as: "Give one possible set of values for the number of times each die number occurs." From this point of view the response 60, 0, 0, 0, 0, 0 is just as valid as any other response where the six numbers sum to 60.

There is also at least one possible interpretation of the question for which "10, 10, 10, 10, 10, 10" is a better response than "12, 11, 9, 11, 8, 9." For example: "I'm going to roll a die 60 times. Try to guess the number of times that each of the numbers 1 to 6 will comes up." Given this scenario, to maximise your chance of "winning" you should choose the mode, which is 10, 10, 10, 10, 10, 10. By the multinomial distribution, your probability of guessing correctly is then

$$\frac{60!}{(10!)^6}\left(\frac{1}{6}\right)^{60} \approx 7.5\times10^{-5}$$

Granted, this is very small probability, but it is a better guess than 12, 11, 9, 11, 8, 9, for which the probability is only

$$\frac{60!}{12!11!9!11!8!9!}\left(\frac{1}{6}\right)^{60} \approx 4.2\times10^{-5}$$

Here is where we risk opening a particularly troublesome can of worms. If we suggest to students that 12, 11, 9, 11, 8, 9 is a good answer to this question, do we risk creating more misunderstandings than we solve? Consider a possible conversation.

Student: You're saying 12, 11, 9, 11, 8, 9 is a better answer than 10, 10, 10, 10, 10, 10. So 12, 11, 9, 11, 8, 9 is more likely to occur than 12, 11, 9, 11, 8, 9?

Teacher: No; it's less likely to occur. But the issue is that 10, 10, 10, 10, 10, 10 is very unlikely to occur, so we shouldn't pick it.

Student: But then 12, 11, 9, 11, 8, 9 is even less likely to occur, so shouldn't we avoid picking it too?

Teacher: But we're more likely to get a pattern like 12, 11, 9, 11, 8, 9 than a pattern like 10, 10, 10, 10, 10, 10.

Student: What do mean by a pattern like "12, 11, 9, 11, 8, 9?" Do you mean we're more likely to get a pattern starting with 12 than a pattern starting with 10?

Teacher: No. It's more likely that we get 10 ones than 12 ones from the 60 throws.

Student: Do you mean we're more likely to get a pattern that contains the number 12 exactly once rather than a pattern with no twelves in it?

Teacher: I'm not sure. I'll need my calculator to check that. It's more that the pattern 12, 11, 9, 11, 8, 9 is representative of lots of patterns with a moderate degree of variability, while 10, 10, 10, 10, 10, 10 is only one pattern.

Student: Now you're changing the question. You asked me to choose just one pattern. You didn't say it could also be representative of other patterns. And why can't my 10, 10, 10, 10, 10, 10 be representative of all the patterns starting with 10?

Our brains have evolved the skill of identifying patterns. Hence we tend to feel that regular patterns like 10, 10, 10, 10, 10, 10 or 9, 9, 10, 10, 11, 11 are somehow more "special" than those like 12, 11, 9, 11, 8, 9. So, we rather illogically view the latter non-special pattern as representative of many such nondescript patterns, and the probability that we obtain a pattern belonging to that large set of nondescript patterns exceeds the probability of getting a single special pattern such as 10, 10, 10, 10, 10, 10.

So we need to be careful that we do not accidentally mislead students into believing a common gambler's fallacy. We could carry out the experiment of rolling the die 60 times many times over and never get a "special" pattern. The common error is to incorrectly conclude from this that any particular "non-special" pattern is thus more likely to occur than any particular "special" pattern. For example, we need to ensure that students realise that in the die roll question the outcome 10, 10, 10, 10, 10, 10 is more likely to occur than 12, 11, 9, 11, 8, 9.

For another example of the fallacy, consider a lottery with 100 000 tickets

numbered 1 to 100 000. Gamblers tend to dislike the early tickets that have very few digits. They argue for example that since they have never seen a two digit number win the lottery, ticket number 27 is less likely to win than ticket 56 320.

These comments are not meant to detract from the aims and methods of Sharma's paper. It is certainly important that students and teachers have a good understanding of the effects of variation, and questions like Item 1 in Sharma's study are excellent tools to this end. We do however need to be careful that discussions of variation effects do not lead to other misunderstandings, such as the gambler's fallacy described above.

Reference

Sharma, S. (2007). Exploring pre-service teachers' understanding of statistical variation: Implications for teaching and research. *Australian Senior Mathematics Journal, 21*(2), 31–43.